

## Report

# Sound Categories Are Represented as Distributed Patterns in the Human Auditory Cortex

Noël Staeren,<sup>1,2</sup> Hanna Renvall,<sup>1,2,3</sup> Federico De Martino,<sup>1,2</sup> Rainer Goebel,<sup>1,2</sup> and Elia Formisano<sup>1,2,\*</sup>

<sup>1</sup>Faculty of Psychology and Neuroscience

Department of Cognitive Neuroscience

<sup>2</sup>Maastricht Brain Imaging Center (M-BIC)

University of Maastricht

6200 MD Maastricht

The Netherlands

## Summary

The ability to recognize sounds allows humans and animals to efficiently detect behaviorally relevant events, even in the absence of visual information. Sound recognition in the human brain has been assumed to proceed through several functionally specialized areas, culminating in cortical modules where category-specific processing is carried out [1–5]. In the present high-resolution fMRI experiment, we challenged this model by using well-controlled natural auditory stimuli and by employing an advanced analysis strategy based on an iterative machine-learning algorithm [6] that allows modeling of spatially distributed, as well as localized, response patterns. Sounds of cats, female singers, acoustic guitars, and tones were controlled for their time-varying spectral characteristics and presented to subjects at three different pitch levels. Sound category information—not detectable with conventional contrast-based methods analysis—could be detected with multivoxel pattern analyses and attributed to spatially distributed areas over the supra-temporal cortices. A more localized pattern was observed for processing of pitch laterally to primary auditory areas. Our findings indicate that distributed neuronal populations within the human auditory cortices, including areas conventionally associated with lower-level auditory processing, entail categorical representations of sounds beyond their physical properties.

## Results

During the fMRI measurements, subjects ( $n = 8$ ) listened to sounds from three “real life” categories (*Singers*, *Cats*, *Guitars*) and synthetic control sounds (*Tones*), presented at three different pitch levels. All “real life” sounds were tonal and had the same fundamental frequency and similar harmonic structure (see [Experimental Procedures](#)). Besides being matched in terms of duration, root-mean-square (RMS) power, and temporal envelope, our stimuli were further manipulated by matching of the temporal profile of their harmonic structure (see [Figure 1](#) and [Audio Files S1–S3](#) [available online]). This novel stimulus manipulation is particularly relevant, because it ensured that the perceptual “pitch” dimension, mainly dependent on the sound fundamental

frequency, was matched across categories. All sounds were delivered binaurally via headphones, in blocks of four, at a comfortable listening level, with the use of a clustered-volume acquisition technique that allowed for presentation of auditory stimuli in silence between subsequent acquisitions (see [Experimental Procedures](#)). Sounds within a block were from the same category and had the same of three possible fundamental frequencies (250 Hz = *Low*, 480 Hz = *Middle*, 920 Hz = *High*), resulting altogether in twelve experimental conditions.

## Univariate Statistical Analysis

Sounds from all of the categories (*Singers*, *Guitars*, *Cats*, *Tones*) evoked significant BOLD responses in a large expanse of the auditory cortex, including bilaterally the Heschl's gyrus (HG), the superior temporal gyrus (STG), and the upper bank of the superior temporal sulcus (STS) (see [Figure S1](#)). With conventional univariate statistical contrasts, consistent differences were detected in the superior temporal regions only for the *Cats versus Tones* comparison in six out of eight subjects at a rather lenient voxel-wise threshold of  $p = 0.01$  (uncorrected; see [Figure S2](#)). Our control on the acoustic sound properties presumably reduced the voxel-by-voxel differences of BOLD responses evoked by the different sound categories.

## Multivariate Pattern Recognition—Learning of Sound “Category”

Contrast-based methods can detect only localized surplus of hemodynamic activity for one condition compared with another, therefore ignoring the potential information of non-maximal responses. As the next step, we therefore used a statistical pattern-recognition approach [6] and tested the hypothesis that the overall spatial patterns of observed responses would convey information on the sound being presented. In each subject, we conducted six pair-wise classification experiments in which sound-evoked response patterns were labeled according to their category (*Singers*, *Cats*, *Guitars*, *Tones*), irrespective of their fundamental frequency. We examined whether our learning algorithm, after being trained with a subset of labeled brain responses (20 trials per category), would accurately classify the remaining unlabeled responses (10 trials per category; see [Supplemental Experimental Procedures](#)).

For all classifications, the recursive algorithm was able to learn the functional relation between the sounds and the corresponding evoked spatial patterns and classify the unlabeled sound-evoked patterns significantly above chance level (0.5), with mean classification correctness across subjects of 0.69 for *Singers versus Guitars* ( $p = 2.8401 \times 10^{-4}$ , two-sided  $t$  test,  $n = 8$ ), 0.69 for *Singers versus Cats* ( $p = 2.5552 \times 10^{-5}$ ), and 0.70 for *Guitars versus Cats* ( $p = 2.6351 \times 10^{-4}$ ) ([Figure 2](#), left).

In order to quantify the consistency of the discriminative maps across subjects, we generated *group-level* maps ([Figure 2](#), right) by cortical realignment [7] of individual discriminative maps. Given that single-subject maps included only voxels that “survived” the recursive elimination of irrelevant features in the algorithm, the group maps can be

\*Correspondence: [e.formisano@psychology.unimaas.nl](mailto:e.formisano@psychology.unimaas.nl)

<sup>3</sup>Present address: Brain Research Unit, Low Temperature Laboratory, Helsinki University of Technology, Finland

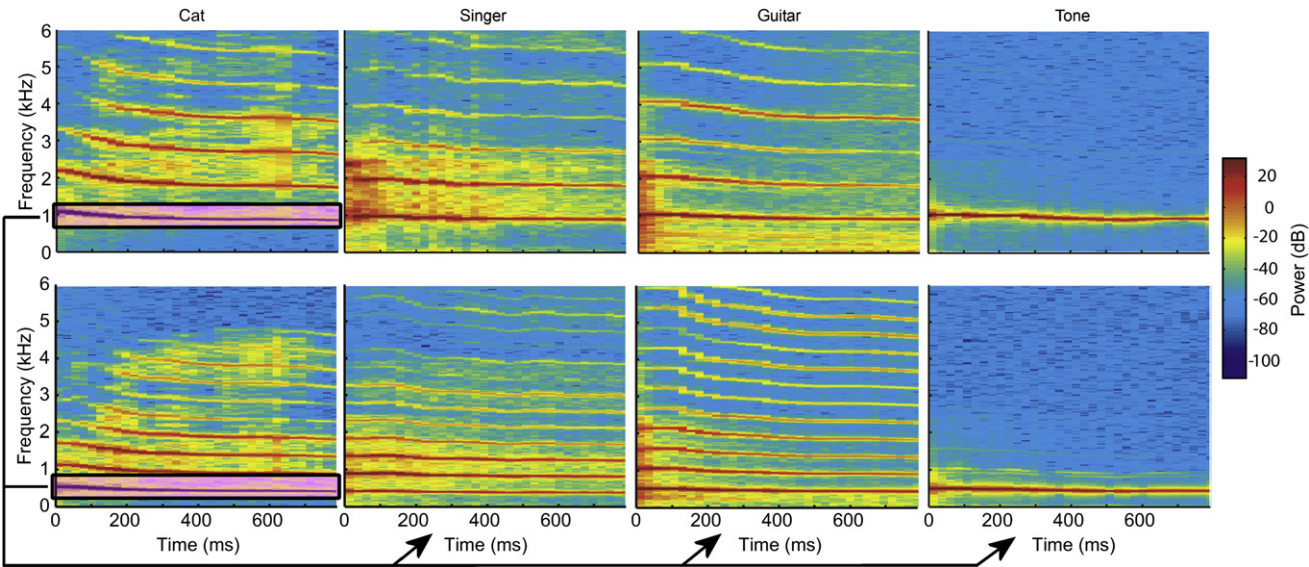


Figure 1. Spectrograms of Exemplary Stimuli  
The four stimulus categories at *High* (920 Hz; top) and *Medium* (480 Hz; bottom) fundamental-frequency levels. The time-varying fundamental frequency of the cat sound (purple rectangle) was imposed onto the other stimuli. The harmonic structure of the sounds was modified accordingly.

interpreted as a representation of spatial patterns that were consistently informative across subjects (see [Supplemental Experimental Procedures](#)).  
At the group level, the distributed activation patterns that differentiated *Singers* from *Guitars* were located at the antero-lateral HG, the planum temporale (PT), and the posterior STG and/or STS in the left hemisphere and at the lateral HG and the middle-posterior STG and/or STS in the right hemisphere. *Singers* were differentiated from *Cats* at the HS, the PT, and the posterior STG in the left hemisphere and at the middle-posterior STG and the PT in the right hemisphere. *Guitars* were differentiated from *Cats* at the left anterolateral HG, the

HS, and the posterior STG and at the right anterolateral HG, the PT, and the middle-posterior STG and/or STS. These results suggest that spatially distributed patterns encoded information on sound category in the superior temporal regions (see [Figure S3](#) for classification accuracies and discriminative maps for differentiating between *Categories* and control *Tones*).

#### Multivariate Pattern Recognition—Learning of Sound “Fundamental Frequency”

Because the stimuli were presented at three different fundamental frequency levels, we conducted a second analysis to

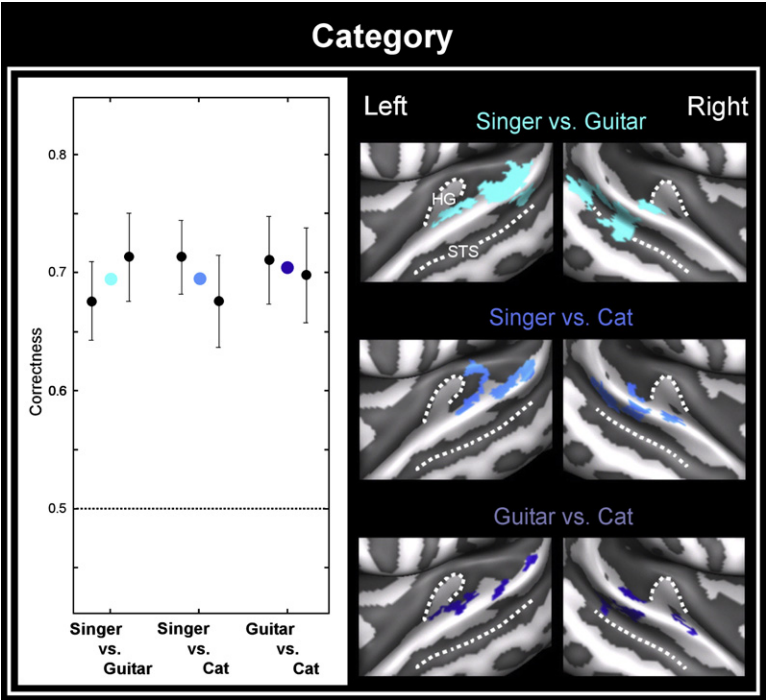


Figure 2. Multivariate Pattern Recognition—Learning of Sound “Category”  
Group-averaged classification accuracies (left) and group discriminative maps (right) for between-category comparisons. For all binary discriminations, the black dots indicate the classification accuracy of test trials for each individual category and the colored dots indicate the classification accuracy averaged over the two categories. Error bars indicate the standard errors. For all classifications, the recursive algorithm was able to learn the functional relation between the sounds and the corresponding evoked spatial patterns and to classify the unlabeled sound-evoked patterns significantly above chance level (0.5). Discriminative patterns are visualized on the inflated representation of the auditory cortex resulting from the realignment of the cortices of the eight participants. A location was color-coded if it was present on the individual maps of at least five of the eight subjects.

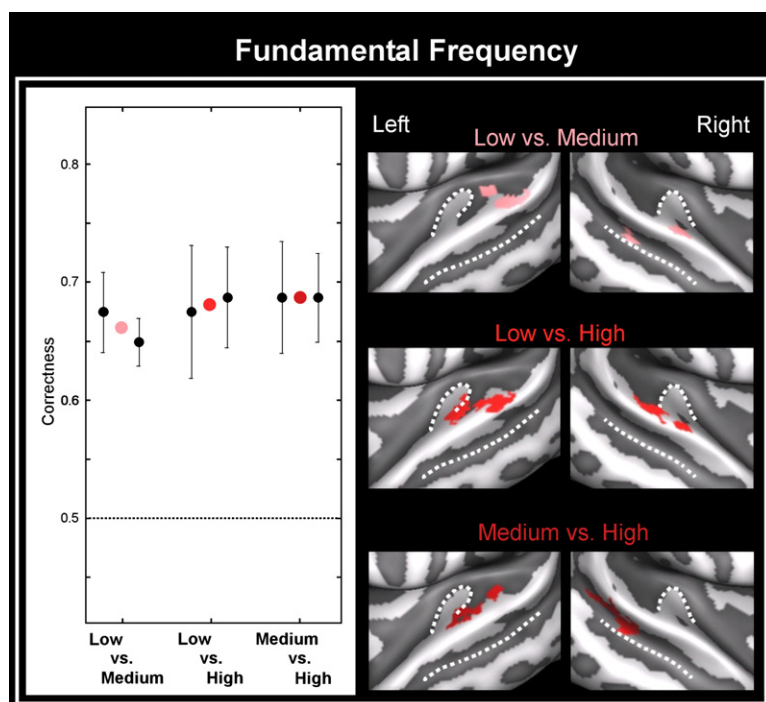


Figure 3. Multivariate Pattern Recognition—Learning of Sound “Fundamental Frequency”

Group-averaged classification accuracies (left) and group discriminative maps (right) for between-frequency comparisons. For all binary discriminations, the black dots indicate the classification accuracy of test trials for each individual frequency and the colored dots indicate the classification accuracy averaged over the two frequencies. Error bars indicate the standard errors. For all classifications, the recursive algorithm was able to learn the functional relation between the sounds and corresponding evoked spatial patterns and to classify the unlabeled sound-evoked patterns significantly above chance level (0.5). Discriminative patterns are visualized on the inflated representation of the auditory cortex resulting from the realignment of the cortices of the eight participants. A location was color-coded if it was present on the individual maps of at least five of the eight subjects.

investigate the regions that were most discriminative with respect to the fundamental frequency. The same sound-evoked response patterns as used in the first analysis were now labeled according to their fundamental frequency (*High*, *Medium*, *Low*), irrespective of their category. The recursive algorithm was then trained to discriminate the fundamental frequencies. All frequency discriminations were statistically significant, with mean classification correctness across

subjects of 0.66 for *Low versus Medium* ( $p = 1.8187 \times 10^{-4}$ , two-sided  $t$  test,  $n = 8$ ), 0.68 for *Low versus High* ( $p = 2.3 \times 10^{-3}$ ), and 0.68 for *Medium versus High* ( $p = 1.224 \times 10^{-4}$ ). Figure 3 shows the resulting group discriminative maps (right) and the corresponding correctness values (left). The group discriminative maps related to

fundamental frequencies were more clustered than the category discriminative maps, and they were circumscribed to the most lateral portion of HG and/or HS bilaterally and to the posterior STG. This finding is in accordance with previous studies indicating the lateral portion of the HG as relevant for pitch processing [8, 9].

Figure 4 summarizes the group discriminative maps obtained for the discrimination of categories (blue) and fundamental frequencies (red). The individual subject maps were in accordance with the group results and are illustrated in Figure S4.

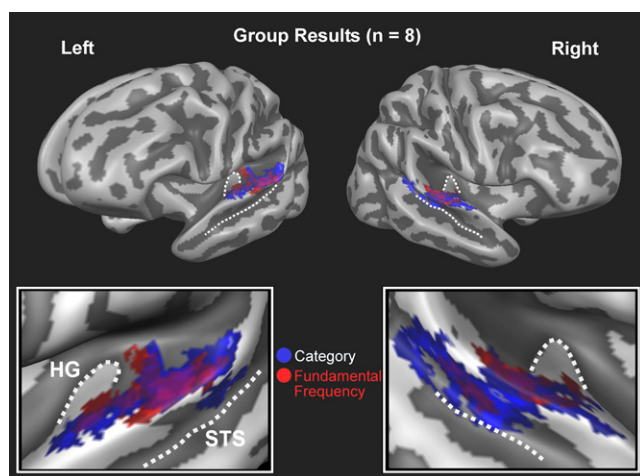


Figure 4. Comparison of Discriminative Maps

The cortex-based aligned group discriminative maps for category (blue) and fundamental frequency (red) discrimination. Category and fundamental frequency discriminative maps were obtained by the logic union of the discriminative maps corresponding to the three binary classifications (Figures 2 and 3, respectively). A vertex was color-coded if it was present on the individual maps of at least five of the eight subjects. This corresponds to a false discovery rate-corrected threshold of  $q = 7.9 \times 10^{-3}$  for the category map and  $q = 2.6 \times 10^{-3}$  for the fundamental frequency map (see Supplemental Experimental Procedures). Note that the discrimination map for fundamental frequency was more clustered than that for category.

## Discussion

### Localized Versus Distributed Representation of Sound Categories

In the present study, we investigated the representation and processing of auditory categories within the human supratemporal cortex. In particular, we asked whether the areas around the primary auditory cortex would code for sound categories irrespective of their physical attributes and, if so, whether these representations would be localized in specialized areas or, rather, distributed across the auditory cortex.

Our investigation differs from previous studies of the “what” auditory-processing stream in terms of stimulus design and data analysis. Because sounds from different categories also tend to differ acoustically, differences in the cortical responses between categories may also reflect merely acoustic-stimulus properties. Use of synthetic sounds would allow a more precise acoustic control [8, 10]; however, natural and synthetic sounds unavoidably differ in terms of ecological validity and familiarity [11, 12]. Here, we used sounds from three “real-life” categories and matched them with respect to many acoustic dimensions, including their duration, average RMS level, amplitude envelope, and harmonic-to-noise ratio [5, 13], as well as the temporal profile of the sound spectrum. Furthermore, by utilizing our recursive method for multivoxel



pattern analysis, we could directly address the issue of localized versus distributed coding of auditory categories in STG and/or STS.

Our results indicate that, similar to the representation of visual-object categories in the ventral temporal cortex [14], representations of sound categories in the superior temporal cortex are widely distributed. Removing most of the physical differences between categories diminished the differences between localized BOLD responses, as reflected by the absence of between-category effects in our univariate analysis. Nevertheless, our iterative multivariate classification analysis showed that the activation patterns could be decoded into categories. Information in the spatially distributed patterns of activity may thus reflect a more abstract perceptual level of representation of sounds.

These findings suggest a revision of previous models that imply a hierarchical processing of auditory categories in the auditory cortices. In these models, the superior temporal cortex is organized in specialized areas among which the neural processing of a sound proceeds from the analysis of its low-level physical constituents to higher perceptual dimensions. Auditory areas with a clear selectivity for a given category, e.g., voice [4, 15, 16], are seen as functional units in which a more abstract representation of a sound is formed. However, the “higher-level” areas show vigorous BOLD responses also to relatively simple stimuli (see the responses to *Tones* in Figure S1), implying sensitivity to “lower-level” sound properties as well. In the present study, the discriminative activation patterns overlapped with—but were not limited to—locations that have been indicated in the previous investigations as functionally specialized areas for human [4] and animal [17] vocalizations. We suggest that a “categorical” representation of a sound emerges from the joint encoding of information occurring not only in a small set of “higher-level” selective areas but also in areas conventionally associated with “lower-level” auditory processing. This suggestion is not without prerequisites: The temporal auditory areas are anatomically heavily interconnected [18], and, even in the “early” auditory areas, neurons exhibit complex dependencies on the auditory input [11, 12]. Furthermore, a distributed cortical coding of sound properties could explain why several auditory regions have been implicated in the processing of many different auditory attributes [19]. The discriminative maps of “category” and “fundamental frequency” overlapped substantially, thus suggesting that regions encoding relatively basic attributes of sounds, such as pitch, or higher level properties, such as category, are not mutually exclusive.

### Univariate Versus Multivariate Modeling of Responses

Machine learning methods allow modeling of distributed patterns of cortical activations. These methods provide better sensitivity than the conventional univariate statistical analyses, by integrating weak but consistent discriminative responses at the single locations and by exploiting the correlations between multiple locations. Thus, these methods can allow the detection of small effects, e.g., those produced by perceptual differences between stimulus categories [20, 21].

We want to point out that in cases in which significant differences between conditions could be detected already at single-voxel level, high classification accuracies were obtained with our multivariate method, and—as expected—the multivariate discriminative maps and the univariate contrast maps overlapped (see Figures S2 and S3). Discriminative maps, however, included additional areas whose joint activity and correlations

were equally informative with respect to the classification of conditions. In the between-category discriminations, accuracy levels were above chance in all of our subjects and were not accompanied by any significant univariate effects. Importantly, corresponding discriminative patterns were highly consistent across subjects.

The minimization of acoustical differences between categories may partly explain why accuracy levels reached in our analyses were lower than those obtained in analogous analyses in the visual domain [14, 22].

With our method, a multivariate analysis does not invariably lead to distributed results. For instance, relabeling of the stimuli on the basis of their fundamental frequency led the same learning algorithm used in the analysis of categories to find substantially different discriminative maps.

### Limitations of Present Stimuli and Extension to Auditory Scenes

The present stimuli were relatively simple: For example, even though our *Singers* stimuli were real voices, their complexity was minimal compared with, e.g., spoken language. Although this resulted in greater stimulus control, it also restricted the spectral richness and ecological validity of our stimuli. Despite our efforts in equalizing low-level acoustic properties, the degree of acoustical similarities between sounds of the same category was higher than between sounds from different categories. It is thus possible that our learning process reflected decoding of a complex combination of spectral and temporal features of our “sound categories.” In future studies, the higher-order representations of natural sounds may also be addressed—for example, by testing the ability of a brain-based classifier to generalize its performance to realistic auditory situations, such as recognizing a voice embedded in a noisy scene after being trained with voices presented in silence.

### Experimental Procedures

#### Subjects

We studied, after obtaining informed consent, eight Dutch and one Belgian undergraduate university students (mean age  $\pm$  SD:  $24 \pm 5$  yrs; eight females and one male; all right-handed). Subjects had no history of hearing or neurological impairments and were naive to the experimental setup. The study received a prior approval by the Ethical Committee of the Faculty of Psychology, University of Maastricht.

#### Auditory Stimuli

The stimuli were 800-ms sounds (sampled at 44.1 kHz) from four sound categories: cats, singing female voices, acoustic guitars, and tones. Each category except the tone category consisted of three different representatives (e.g., three different singers). For the addition of acoustical variability to the stimuli, all sounds were transposed to three different fundamental frequencies (250, 480, and 920 Hz), thus resulting altogether in twelve conditions. The values of fundamental frequencies were chosen so as to ensure that stimuli were clearly recognizable and to avoid pure octave pitch differences (e.g., 250, 500, and 1000 Hz).

For equalization of the spectrotemporal profiles and the perceptual pitch of the stimuli, the time-varying fundamental frequency of the cat sounds was extracted on 25 time points within each stimuli, with Praat software [13]. These pitch profiles were then used as references for pitch bending in Adobe Audition and applied to all other sounds. Note that not only was the fundamental frequency of the manipulated sounds adjusted, but all related harmonics (see Figure 1 and Audio Files S1–S3) were adjusted as well. For the selected sound categories, continuous pitch changes are natural (e.g., sliding in between two tones when singing, or bending a guitar string). Tones were used as control sounds.

The sounds were low-pass filtered at 14 kHz for five subjects and, for further minimization of the acoustical differences between sound

categories, at 7 kHz for three subjects. No significant differences between the results of these groups were found in the statistical analyses, and subjects were thus grouped together in the reported results. The sound amplitude envelopes and average root-mean-square levels were matched with MATLAB 7.0.1 (MathWorks, Natick, MA, USA). The harmonic-to-noise ratio [5, 13] was significantly different only between tones and sound categories ( $p < 0.001$ ), not between categories ( $p > 0.05$ ).

#### Experimental Paradigm

Brain imaging was performed with a 3 Tesla Siemens Allegra (head setup) at the Maastricht Brain Imaging Center. During the measurements, the stimuli were delivered binaurally via magnetic resonance-compatible headphones (Commander XG, Resonance Technology, Northridge, CA) in blocks of four at a comfortable listening level. For minimizing the effect of scanner noise, the sounds were presented during 1600-ms silent periods between 2000-ms scans; the 800-ms sounds were preceded and followed by a 400-ms silence, with the use of a clustered volume EPI technique (23 slices covering the perisylvian cortex; see [Supplemental Experimental Procedures](#)) that allowed for presentation of auditory stimuli in silence between subsequent volume acquisitions [23–25]. The stimuli within a block were from the same category and frequency level, resulting in altogether twelve experimental conditions. The experimental blocks had a duration of 14.4 s. The conditions were repeated in a pseudorandom order and were followed by a rest period of identical length, at the beginning of which the subjects were asked to respond with a button press to indicate whether the last two sounds in the block were the same (50% of the trials). The response hand was alternated across subjects.

See [Supplemental Data](#) for details on stimulus training, scanning parameters, preprocessing, cortex-based alignment procedure, univariate data analysis, and multivariate pattern recognition.

#### Supplemental Data

Supplemental Data include Supplemental Experimental Procedures, four figures, one table, and four audio files and can be found with this article online at [http://www.current-biology.com/supplemental/S0960-9822\(09\)00740-4](http://www.current-biology.com/supplemental/S0960-9822(09)00740-4).

#### Acknowledgments

We would like to thank Lars Riecke for his useful comments on the experimental design and on the manuscript. Funding for the present research was contributed to E.F. (Vernieuwingsimpuls VIDI, 452-04-330) from the Netherlands Organization for Scientific Research (NWO) and to H.R. from the Academy of Finland.

Received: July 11, 2008

Revised: December 12, 2008

Accepted: January 27, 2009

Published online: March 5, 2009

#### References

1. Kaas, J.H., and Hackett, T.A. (1999). 'What' and 'where' processing in auditory cortex. *Nat. Neurosci.* 2, 1045–1047.
2. Rauschecker, J.P., and Tian, B. (2000). Mechanisms and streams for processing of "what" and "where" in auditory cortex. *Proc. Natl. Acad. Sci. USA* 97, 11800–11806.
3. Adriani, M., Maeder, P., Meuli, R., Thiran, A.B., Frischknecht, R., Villemure, J.G., Mayer, J., Annoni, J.M., Bogousslavsky, J., Fornari, E., et al. (2003). Sound recognition and localization in man: specialized cortical networks and effects of acute circumscribed lesions. *Exp. Brain Res.* 153, 591–604.
4. Belin, P., Zatorre, R.J., Lafaille, P., Ahad, P., and Pike, B. (2000). Voice-selective areas in human auditory cortex. *Nature* 403, 309–312.
5. Lewis, J.W., Brefczynski, J.A., Phinney, R.E., Janik, J.J., and DeYoe, E.A. (2005). Distinct cortical pathways for processing tool versus animal sounds. *J. Neurosci.* 25, 5148–5158.
6. De Martino, F., Valente, G., Staeren, N., Ashburner, J., Goebel, R., and Formisano, E. (2008). Combining multivariate voxel selection and support vector machines for mapping and classification of fMRI spatial patterns. *Neuroimage* 43, 44–58.
7. Goebel, R., Esposito, F., and Formisano, E. (2006). Analysis of functional image analysis contest (FIAC) data with brainvoyager QX: From single-subject to cortically aligned group general linear model analysis and self-organizing group independent component analysis. *Hum. Brain Mapp.* 27, 392–401.
8. Patterson, R.D., Uppenkamp, S., Johnsrude, I.S., and Griffiths, T.D. (2002). The processing of temporal pitch and melody information in auditory cortex. *Neuron* 36, 767–776.
9. Griffiths, T.D. (2003). Functional imaging of pitch analysis. *Ann. N.Y. Acad. Sci.* 999, 40–49.
10. Warren, J.D., Jennings, A.R., and Griffiths, T.D. (2005). Analysis of the spectral envelope of sounds by the human brain. *Neuroimage* 24, 1052–1057.
11. Nelken, I. (2004). Processing of complex stimuli and natural scenes in the auditory cortex. *Curr. Opin. Neurobiol.* 14, 474–480.
12. Wang, X., Lu, T., Snider, R.K., and Liang, L. (2005). Sustained firing in auditory cortex evoked by preferred stimuli. *Nature* 435, 341–346.
13. Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot Int.* 5, 341–345.
14. Haxby, J.V., Gobbini, M.I., Furey, M.L., Ishai, A., Schouten, J.L., and Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science* 293, 2425–2430.
15. Warren, J.D., Scott, S.K., Price, C.J., and Griffiths, T.D. (2006). Human brain mechanisms for the early analysis of voices. *Neuroimage* 31, 1389–1397.
16. Belin, P., Fecteau, S., and Bedard, C. (2004). Thinking the voice: neural correlates of voice perception. *Trends Cogn. Sci.* 8, 129–135.
17. Petkov, C.I., Kayser, C., Steudel, T., Whittingstall, K., Augath, M., and Logothetis, N.K. (2008). A voice region in the monkey brain. *Nat. Neurosci.* 11, 367–374.
18. Tardif, E., and Clarke, S. (2001). Intrinsic connectivity of human auditory areas: a tracing study with Dil. *Eur. J. Neurosci.* 13, 1045–1050.
19. Formisano, E., De Martino, F., Bonte, M., and Goebel, R. (2008). "Who" is saying "what"? Brain-based decoding of human voice and speech. *Science* 332, 970–973.
20. Kamitani, Y., and Tong, F. (2005). Decoding the visual and subjective contents of the human brain. *Nat. Neurosci.* 8, 679–685.
21. Haynes, J.D., and Rees, G. (2005). Predicting the orientation of invisible stimuli from activity in human primary visual cortex. *Nat. Neurosci.* 8, 686–691.
22. Cox, D.D., and Savoy, R.L. (2003). Functional magnetic resonance imaging (fMRI) "brain reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex. *Neuroimage* 19, 261–270.
23. Riecke, L., Van Opstal, A.J., Goebel, R., and Formisano, E. (2007). Hearing Illusory Sounds in Noise: Sensory-perceptual Transformations in Auditory Cortex. *J. Neurosci.* 27, 12684–12689.
24. Jancke, L., Wustenberg, T., Scheich, H., and Heinze, H.J. (2002). Phonetic perception and the temporal cortex. *Neuroimage* 15, 733–746.
25. van Atteveldt, N., Formisano, E., Goebel, R., and Blomert, L. (2004). Integration of letters and speech sounds in the human brain. *Neuron* 43, 271–282.